LA-UR-20-20586
Approved for public release; distribution is unlimited.

Title:	In-Situ Inference: Bringing Advanced Data Science Into Exascale Simulations
Author(s):	Urban, Nathan Mark Lawrence, Earl Christopher Biswas, Ayan
Intended for:	Sharing with potential collaborators.
Issued:	2020-01-21

In-Situ Inference

PI: Urban, Nathan M; CCS-2; nurban@lanl.gov

Project Goals

"What new science would be possible if you had ALL the data?"

As simulations generate ever-increasing amounts of data, there are correspondingly richer opportunities for analysis and scientific discovery—discoveries that will be missed if most of the data must be discarded before it is analyzed. Because future exascale architectures will be increasingly storage-limited, it will not be possible to save the vast majority of simulation data for later analysis, requiring analysis to occur "*in-situ*" within the simulation. However, existing in-situ data analysis frameworks provide little or no support for one of the most sophisticated forms of data science: probabilistic statistical modeling or uncertainty quantification (UQ), and the accompanying challenge of *inference*—fitting those statistical models to massive simulation output. Our goal is to develop the fundamental statistical algorithms and computer science needed to perform *statistical inference in-situ* (in HPC simulations) to the full stream of data those simulations generate.

Consider the mission science challenge of quantifying the probability of events in predictive HPC simulations, and understanding the underlying factors influencing the likelihood of these events. Examples include the future risk of extreme weather events damaging population centers, or of extreme electron flux events in solar storms damaging satellites.

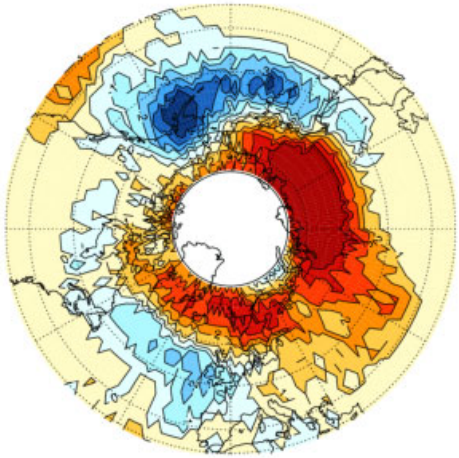

Figure 1. Phenomena in simulation data, such as atmospheric blocking events leading to blizzards and cold snaps,[56] exhibit complex patterns of spatiotemporal variability.[7] Sophisticated statistical inference is required to identify relationships between such events and changing environmental conditions. But it is currently impossible either to store the data needed to fit such a statistical model offline, or to fit it online in the simulation.

To understand why this is a *statistical inference* challenge, turn to questions that we cannot yet adequately answer: How will the frequency of blizzards change as the climate warms (Fig. 1)? How much of this change is attributable to sea ice retreat, vs. surface warming, vs. enhanced moisture transport? How do the statistics of turbulent plasma flows change as a function of solar cycle or prior history of the magnetospheric state (Fig. 2)? The tools needed to answer these questions are *statistical models*: probability density estimation, extreme value analysis, nonstationary spatial and time series modeling, regression and covariance analysis to quantify the sensitivity of effects to causes, etc. The *inference algorithms* used to fit these statistical models to data include Bayesian inference and Monte Carlo sampling, but they currently only work offline, on highly-reduced data.

The above grand challenge questions, by contrast, require *all* the data to answer. We are asking statistical questions down to the individual grid cell and near-timestep level in exascale simulations, looking for subtle statistical differences in probability distributions at different locations and times, and the dependence of event frequencies on vaguely-defined phenomena that extend throughout a vast 3D domain (such as mesoscale weather formations or geomagnetic substorm injections). In such settings, and when we are looking to quantify potential dependencies of any data point with any other, we cannot simply identify all the relevant features of interest ahead of time. Without new algorithms and computer science to *infer* or fit sophisticated statistical models *in-situ*, to *all* of the simulation data as it is being generated, modern data science will be left behind in the exascale revolution.

Background and Statement of Problem

There is an increasing disconnect between state-of-the-art statistical inference applications and the analyses that scientists currently can perform. This disconnect exists because (1) HPC I/O limitations force analysis to occur in-situ; (2) in-situ data reduction loses too much information for inference, and forces inference algorithms to be designed around the needs of data reduction, rather than the needs of science; and (3) existing inference methods primarily work offline. Bottlenecks to in-situ inference are:

(B1) inference methods are not scalable for HPC
(B2) software supporting in-situ programming is cumbersome and inefficient for advanced data science

As a result, even simple in-situ statistical analyses are major programmatic activities requiring teams of data scientists, computer scientists, and model developers working to implement complex, model-specific pipelines of data reduction and inference, rather than something a single data scientist could code and insert in a model.

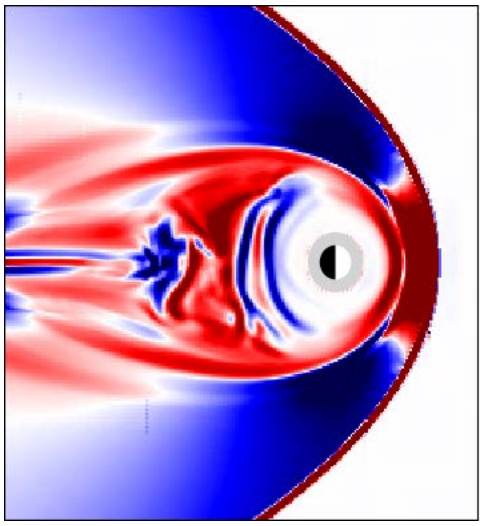

Figure 2. Bursty bulk flows (BBFs) of plasma in the Earth's magnetosphere can drive strong electric currents that damage satellites, simulated here by the SHIELDS model.[50] Understanding the hazard to space operations depends on quantifying the likelihood of intense BBFs, their relationship to large-scale flows, and their dependence on the upstream solar wind.

Examples of potential inference applications include (1) UQ for complex multiscale data, identifying probabilistic relationships between phenomena across scales that would be lost in any plausible data reduction scheme; (2) generative modeling or model reduction, where the goal is to synthesize new examples of the system response that statistically "behave like" the simulation with high fidelity (e.g. stochastic weather generators,[87,47] turbulence emulators[93]); (3) intelligent sampling or steering of simulations towards interesting behavior. Existing in-situ data reduction approaches do not produce the information needed to perform this kind of inference. They focus on computing statistics like means, variances, and correlations; feature extraction, like turbulent eddies and dark matter halos;[91,41,58] and visualization.[1,26,27,28] Any further postprocessing, such as statistical modeling, occurs offline on these reduced data features.[40,47]

This presents two problems for statistical inference. The first is *inference quality*: It assumes that nothing is lost during the in-situ data reduction step. However, as we argued above, this premise is untenable: as the scientific questions we ask become more complex, statistical methods will need to model increasingly subtle relationships within ever-larger data sets. The second problem is *scientific productivity*: This paradigm forces labor-intensive workflows requiring one-off, application- and model-specific coding. The need to first reduce the data, then separately infer what is needed from only those reduced features, requires data scientists to develop entirely different models and inference procedures other than those best suited to the data.

To date, advanced data science algorithms like spatiotemporal and/or Bayesian models are often not scalable to DOE leadership computing data. Even in flagship DOE codes, some of the simplest algorithms, such as principal component analysis (PCA),[89] are sufficiently onerous to modelers that they have never been implemented in-situ; and there is a vast gap in complexity between PCA and the data science applications discussed earlier. We are aware of only one petascale Bayesian inference application,[76] for offline astronomical imagery. The popular Gaussian process (GP)

model of spatiotemporal data is inefficient in its original form, requiring a host of modifications to be more scalable.[65] But common distributed and streaming (online) variants of these statistical algorithms, needed to achieve HPC performance, assume data points are independent or else can be accessed in any order or loaded on any node desired.[35,25,39,71] This does not apply to physical simulation output, which is sequential in time, and data at fixed locations in the physical domain always live on the same nodes (spatial partitioning) determined by the HPC code. Since simulation data are inherently spatiotemporal, and the primary obstacle to "embarrassingly parallel" inference is correlations in space and time, we are missing the fundamental building block of more general inferential analysis in HPC codes: scalable, distributed, streaming spatiotemporal inference.

There are currently no high-level programming environments for in-situ analysts to write general, arbitrary, high-performance inferential algorithms. High-level languages would provide much greater productivity than doing data science in Fortran/C++. A number of in-situ software infrastructures have been developed,[4,30,90,67,88,1] but most are oriented towards visualization or low-level tasks like I/O management. Python is an increasingly popular option for scientific programming, but is not natively efficient on numerical HPC problems. Custom frameworks like Numpy and TensorFlow have emerged to fill this gap, but are essentially separate programming languages implemented in C++ and wrapped in Python. This forces programmers to work within a highly restricted subset of features provided by these libraries to achieve performance, preventing these frameworks from automatically interoperating, and making it difficult to take full advantage of Python's data science ecosystem.[73] Another challenge is that current compilers often miss large performance and portability opportunities for parallel codes,[80,81] preventing data scientists from writing general efficient code without enlisting labor-intensive software engineering assistance from computer scientists.

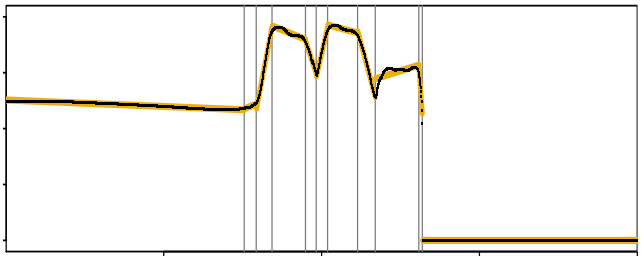

Figure 3. Piecewise linear regression is a simple inference application that can be implemented in-situ to compactly model HPC simulation time series data.

Preliminary studies

Our team has a strong track record in the building blocks necessary for in-situ inference. In addition to general expertise in spatiotemporal modeling,[36,61,38] we have developed an in-situ inference framework for compactly modeling the time variation of features in simulation data using a simple statistical model, piecewise linear regression (Fig. 3).[68] We have also developed sparse Gaussian process methods to fit to offline simulation output (Fig. 4),[55] although they are not yet fully scalable for HPC. We have developed an approach to using high-resolution timestep-level sampling of simulation output to construct a dynamical reduced model of the simulation.[24] In E3SM we have implemented in-situ feature extraction for eddy detection.[91] We have developed intelligent data sampling algorithms under the in-situ ALPINE project.[60,10,11] For modeling the data and performing in situ distribution-driven data summarization we have also explored Gaussian mixture model (GMM) based in-situ inference techniques that have produced promising results.[26,28] The distribution-based data summaries were used successfully to find important features in the data set, track these features, and also to reconstruct the full-resolution data for visual exploration. We have also explored in-situ predictive feature detection using fuzzy-rule based systems where the training was done off-line and the prediction for ensemble simulation runs was performed in situ.[27] We have also recently demonstrated advances in using concurrency-aware compilation techniques to speed up parallel codes.[81,54]

Proposed Innovation

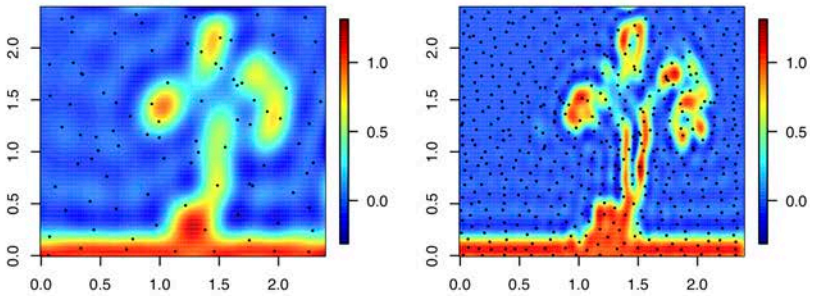

Figure 4. Sparse Gaussian processes are scalable statistical models of massive spatiotemporal data we will fit *in-situ* during simulation. Users trade fast approximation (left) for inference quality (right) by increasing the statistical model complexity.

Scientists have been forced to change their analysis methods to accomodate the limitations of in-situ feature extraction, which is both scientifically inadequate, due to loss of information, and very labor-intensive. We propose another path, which approaches the data science ideal: Allow data scientists to write down *the model they need*, and let them fit it directly, to *the data they need*, without discarding information. This enables new science applications, requiring full interactive access to simulation data, and greatly simplifies analysis and methods, with no need to deal with reduced or missing data.

To realize this vision, we will create fundamental building blocks of *in-situ inference* appropriate for exascale physical simulations: generalizable *spatiotemporal statistical models* and fast *Bayesian inference* algorithms. To work in-situ, the algorithms must be scalable, distributed, streaming (since simulation state variables are overwritten each timestep), and compatible with common HPC model data layout and internode communication patterns. Fast approximations may be required in order to perform inference without significant impact on the simulation runtime.

(Innovation 1) Since physical simulations generate spatiotemporal data, the workhorse in-situ statistical model we will create is a Bayesian probabilistic spatiotemporal model. We will develop a deep sparse Gaussian process (SGP) model of spatiotemporal data,[5,23] with linear-complexity computational scaling properties,[18] and adapt it for HPC use by (1) modifying distributed SGP algorithms to work with spatially-partitioned data from HPC domain decompositions, (2) modifying streaming SGP algorithms for temporally autocorrelated data, and (3) developing new deep learning methods to replace expensive correlation calculations between off-node data with compressed features of the data optimal for the inference problem being solved. These innovations will solve many of the technical issues that arise in more general in-situ inference problems, because distributed and streaming computations are particularly challenging in the HPC setting due to the spacetime correlations amongst data that our spatiotemporal model is designed to handle.

Our spatiotemporal model can be fit directly to simulation output, or can be u a building block in a more complex inference pipeline: a *hierarchical Bayesian spatiotemporal model*.[3] We will do the latter, building a statistical model for the entire probability distribution of a simulation variable, where the parameters of the probability distribution themselves vary smoothly over space and time according to our spatiotemporal model. We will use variational inference (VI), an optimization-based approximate inference algorithm 100–1000× faster than Markov chain Monte Carlo[12,57,72] and highly concurrent. We will exploit the natural analytic variational formulation of our spatiotemporal model to accelerate inference, leaving the development of fully general VI to future work.

(Innovation 2) We will increase productivity by enabling scientists to write high-level, performant numerical algorithms in the Julia scientific computing and data science language,[9] providing seamless abstractions to interweave numerical computing,[74,8] ML,[45,44,75] concurrency,[17] and GPU execution[6,77,31] at HPC scale.[76] We will provide lightweight hooks, data-access abstractions, and shared memory concurrency primitives to execute parallel Julia code coupled to simulations,

hiding implementation details from in-situ programmers. The result will be to provide a higher level of abstraction between data science and HPC implementation to allow data scientists to focus more on statistical modeling and inference, reducing time and labor to implement new forms of analyses and getting them to perform well on new architectures.

Technical impact

The technologies developed here would dramatically advance in-situ data science capabilities. Right now most in-situ analyses are essentially hand-crafted. Higher-level statistical modeling and inference libraries, designed to be scalable and portable across HPC codes, will allow data and domain scientists to focus on modeling the problems they want to solve, instead of on implementation details, greatly accelerating scientific productivity.

Mission impact

Large-scale simulations span all LANL mission areas in simulation and computation, including environmental and space science, nuclear science, materials science, etc. They will increasingly produce an unprecedented amount of data, but only a fraction can be analyzed offline. In-situ inference will permit full analysis of simulation data with the most advanced statistical and ML algorithms available. Arguably an in-situ inference capability will be required for almost any exascale science that requires fitting sophisticated statistical models to simulation data.

R&D Methods and Anticipated Results

We will focus our science applications on a general class of grand challenge problems involving the thorough statistical characterization of spatiotemporal events in simulation data. To demonstrate generality, we will apply our methods to two different HPC codes, E3SM and SHIELDS.

E3SM is a new Earth System Model that was developed to focus on science questions relevant to the DOE mission including risks to water and energy security, which are dependent on many processes (e.g., hurricanes, thunderstorms, atmospheric blocking, atmospheric rivers) that span many orders of magnitude in spatial and temporal scales. Most of these processes require high resolution, less than 25 km in the horizontal, and hours to days in time over a century-scale simulation. With E3SM we will focus on characterizing the risk of mid-latitude winter weather extremes (blizzards and polar-vortex cold snaps)[20,92,49] and how it varies over space and time, as a function of changes to sea ice and other large-scale modes of climate variability.[79,53]

The SHIELDS framework[50] represents an end-to-end model of the Earth's magnetosphere[86] driven by the dynamic solar wind. We will characterize the likelihood of intense small-scale flow structures called bursty bulk flows (BBFs) that can produce strong electric fields hazardous to power and gas lines, and can energize charged particles that damage satellites.[51] To capture the small-scale flow channels and understand their spatiotemporal correlations, the near magnetotail must be resolved at the scale of a few hundred km, and BBF evolution over seconds to minutes. However, strong BBFs do not constantly occur, and their characteristics and occurrence depend nonlinearly on solar wind driving. Therefore many geomagnetic storms, lasting several days each, must be simulated to capture the parameter space of current densities and flow velocities in BBFs.

Methods

Statistical model. Let $Y(x,t)$ be a spacetime field of interest, such as precipitation or electron flux. The inference problem is to estimate the conditional probability distribution $p(Y(x,t)|Z)$, where Z is a high-dimensional vector of predictor variables, such as the full 3D atmospheric or geomagnetic state influencing Y. The probability distribution can be non-Gaussian, different at every location

and time, and dependent on the simulation state Z. The statistical model allows us to ask questions about the probability of, say, a 20" snowfall event (Y) anywhere on the globe at any time, and quantify how sensitive that probability is to changes in surface and stratospheric temperature, sea ice extent, the phase of global climate patterns like the Arctic Oscillation, etc. (Z). Extreme fluctuations in Y will be captured by a generalized extreme value distribution $GEV(Y;\varphi)$ whose parameters φ quantify properties like the expected frequency or return period of rare events. Importantly, estimating the GEV parameters at each location in space *independently* would lead to very noisy estimates, since rare events give rise to small sample sizes. We will regularize the GEV parameter estimates by smoothing them in space and time using a Gaussian process spatiotemporal model, assuming that nearby points have highly correlated extreme parameters.[21] The GEV parameters become *latent spatiotemporal fields*, $\{\varphi(x,t)|Z\}$: fields of unobservable, uncertain statistical parameters rather than simulation variables like Y (Fig. 5).[33] Fitting a spatiotemporally-varying extreme value distribution is important in its own right for our application, but is also a proxy for more complex probabilistic estimation problems. We will use a similar spatiotemporal smoothing approach to model the non-extreme or "central" part of the distribution, as a mixture of Student-t distributions (to capture heavy tails).[70]

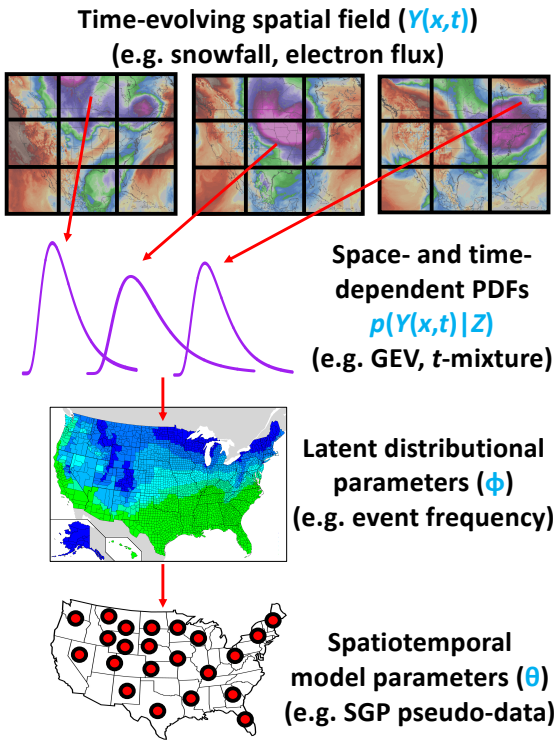

Figure 5. The main statistical model we will develop will characterize the probability density function (PDF) of events in a simulation, how that PDF varies in space and time, and its dependence on large-scale features of the simulation data.

Spatiotemporal model. The core of our statistical model is a probabilistic model of spatiotemporal data. We are not fitting the spatiotemporal simulation variables directly (Y), but rather we are *inferring* the latent distributional parameters φ discussed above, which vary in space and time. For spatiotemporal modeling we will use a more expressive "deep" version[23] of the sparse Gaussian process (SGP); Figs. 4,6.[5,18,85,82] An SGP models a high dimensional, potentially noisy data set of size N using a smaller number ($M \ll N$) of "artificial" data points, or "pseudo-data", greatly improving the scalability of Gaussian processes since correlation-related calculations only need to take place between relatively few pseudo-data points, instead of the full data. The full data are approximated by a smooth interpolation of the pseudo-data, assuming nearby points are correlated with each other according to a

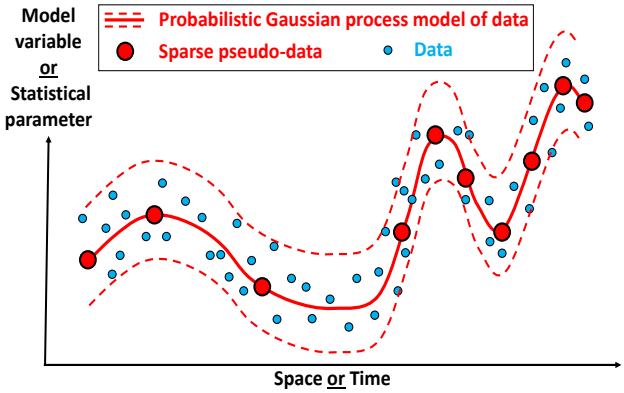

Figure 6. A sparse Gaussian process (SGP; red curve) compactly represents spatial or temporal data (blue dots) by interpolating a small number of representative "pseudo-data" points (red dots) learned from data. SGPs maintain a probabilistic representation of data uncertainty (dashed error bars) and correlations. We will use SGPs to infer smooth variations in unknown quantities (red curve), like the return period of extreme events, from noisy statistical estimates (blue dots).

covariance function fit to data. Pseudo-data are not a subset of the full data; they are *inferred* from a variational optimization procedure to minimize reconstruction error of the original data.

Scaling the statistical model. Further scalability will be achieved by a linear-time variant of SGP[18] and a fast iterative Krylov method.[84] However, SGP still involves inferring the potentially millions of parameters (mostly pseudo-data), which requires distributed computation. Furthermore, correlations across multiple timesteps cannot be directly calculated since state variables are overwritten every timestep. Our approach to both of these challenges is to derive low-dimensional summary statistics that approximate the information needed for inference that is contained in off-node data (to minimize distributed communication) or past-time data.

In the distributed setting we will develop a version of "global-local" inference[83,34] (Fig. 7): each node will fit independent SGPs to the data living on-node, which initially will not lead to a globally consistent inference.[19] We will use SGP pseudo-data as summary statistics, and pass a subset of them between nodes as stand-ins for the full off-node data when performing inference. Each node will then re-fit its data using this additional global information. For streaming inference, it will not be practical to retain in memory the pseudo-data at all previous timesteps. We will apply the streaming sparse Gaussian process (SSGP),[15] which uses a *single*, carefully constructed set of pseudo-data to summarize information about the posterior distribution (not the data itself) contained in all previous timesteps. However, SSGP assumes each spatial snapshot of data is independent in time. We will modify SSGP to accommodate a first-order Markov process allowing successive snapshots to be correlated in time (with a possible linear trend), similar to a Kalman filter assumption.[37]

Scalable Bayesian inference. To infer the parameters of the GEV and central distribution models,[33] we will use variational inference (VI), an approach to Bayesian inference in which the true posterior distribution is approximated within a specified class of distributions.[12] Inference becomes an optimization to find the best fit within the class. In distributed inference we communicate SGP pseudo-data between nodes, but these data are not necessarily the minimal *sufficient statistics* needed to perform inference, leading to unnecessary internode communication. We will apply deep neural networks (DNNs) to learn smaller sufficient statistics.[14,48,16,2] VI optimizes statistical model parameters by maximizing an objective function, the evidence lower bound (ELBO), via iterative gradient ascent. We will train a pair of neural networks to directly predict the gradient that a node needs to update its own parameters. A *summarization* network will compress off-node data, to be globally communicated between nodes, and a *gradient prediction* network will estimate the ELBO gradient for VI, as a

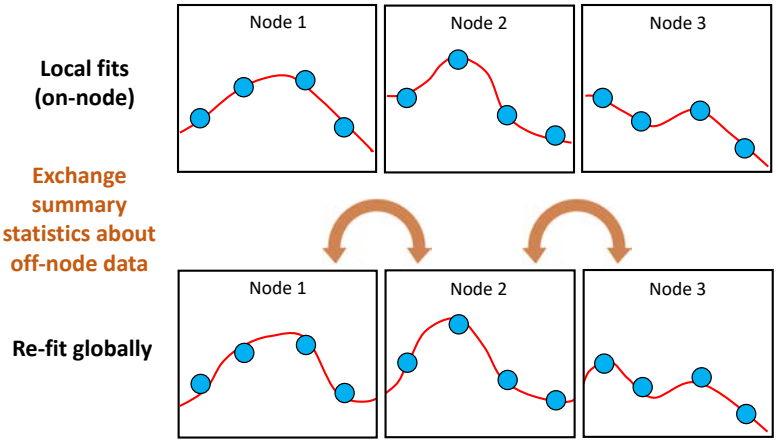

Figure 7. We will take a 3-step approach to distributed inference: (1) Each node fits a local statistical model to all of its own data, neglecting correlations with off-node data, which may lead to discontinuous fits across nodes. (2) Each node computes *summary statistics* of its own data, such as sparse Gaussian process pseudo-data or ML sufficient statistics, and exchanges these low-dimensional summaries with other nodes at low communication overhead. (3) Each node adjusts its inference using off-node summary statistics to produce a globally-consistent fit.

function of on-node data and the off-node data summarized by the first network.[42,43,64,66] The DNNs will be trained to reproduce the true ELBO gradients from synthetic training data simulated offline from a variety of representative SGP models,[43] at a variety of simulated resolutions, [13 rather than always the highest (HPC code) resolution, to accelerate training.

The computational foundations of in-situ science. Our goal is to allow in-situ programmers to write ordinary analysis code without, to the extent possible, needing to directly interact with an HPC code. Inspired by Ascent,[59] the ALPINE many-core in-situ analysis infrastructure, we will develop a generic abstraction layer to two-way couple the Julia language runtime to Fortran/C++ HPC codes (Fig. 8), with specialized data access interface layers that can be written to traverse code-specific data structures and memory layouts, e.g., unstructured mesh data. Our lightweight in situ abstraction layer will provide simple state variable and mesh retrieval functions across these simulations by communicating with the model couplers directly, returning zero-copy data. Using our scheme, simulation data will appear to the in-situ programmer as normal Julia objects. A lightweight data processing workflow will avoid expensive data copies by directly manipulating simulation data pointers, making it easy to couple with HPC codes. We will implement a "blocking" in-situ system where simulations pause to run in-situ analysis, but will also explore asynchronous operation where in-situ analysis runs from a non-blocking call in parallel with the main execution.

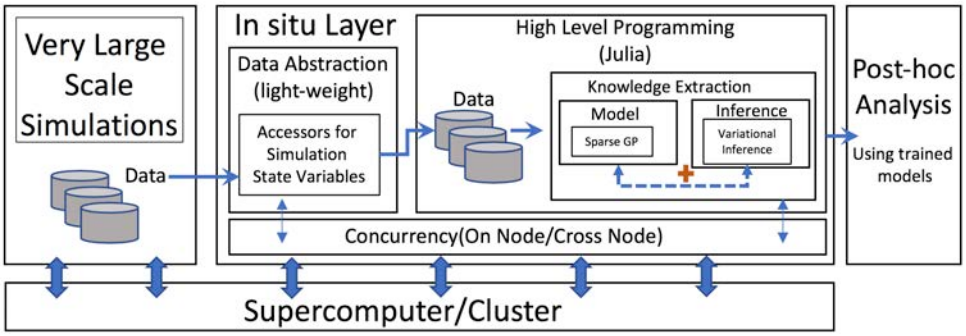

Figure 8. In-situ inference will sit alongside the HPC simulation through a light-weight data abstraction layer, sharing model state variables with a statistical model implemented on top of a variational inference engine. The output of the in-situ inference will be a trained statistical or ML model.

For between-node, or distributed, concurrency we use Julia's abstraction capabilities and MPI wrappers to construct data types that appear as a single shared-memory object to the user, managing MPI distributed communication "under the hood". One bottleneck in the in-situ inference is the computation of SGP covariance matrices and their inverses. We will use iterative Krylov methods to compute fast linear solves with a high degree of parallelism.[29,22,63]

To address the obstacle for in-situ programmers created by the lack of compiler support for parallel optimizations and portability, we will extend LANL's LLVM-based Kitsune project[54] to take advantage of advanced parallel-aware compiler optimizations. Because Julia uses LLVM, we will implement a frontend for the shared memory concurrency primitives in Julia's standard library,[52,17] compiling to Kitsune's parallel intermediate representation. This ensures portability of data science codes by targeting multiple backends, including OpenMP, allowing the HPC simulation's shared memory concurrency to interoperate with the in-situ analyses. This is an important consideration when composing multiple software components on a node: if they use different runtimes, those runtimes can compete for resources, hurting performance. Kitsune's ability to target multiple runtimes ensures that we target the same runtime the application uses, allowing for fine-grained, low-overhead synchronization between the in-situ analysis and the application. In addition to better portability, we expect the parallel-aware optimizations will provide performance improvements not available to standard vendor compilers. Examples of optimizations Kitsune is able

to perform that standard compilers are not include parallel loop-invariant code motion, sync elimination, and constant propagation to concurrent code, leading to large performance improvements.[80,81] When combined with an LLVM-based Fortran compiler,[32] this opens the possibility of optimizations that cross between in-situ analysis and the HPC code, and flexibility in runtime choice, both of which can greatly increase performance.[81]

Expected Results

The deliverables are (1) two fitted probabilistic models of the distribution of extreme and non-extreme events and the dependence of that distribution on space, time, and external predictor variables; (2) scalable, distributed, streaming inference algorithms to fit these models (including a new scalable spatiotemporal model); (3) a high-level programming framework for in-situ data science; (4) prototype code implementing (1)-(3) within the E3SM and SHIELDS simulation codes; and (5) statistical analyses of blizzards and bursty bulk flows using (1). The methods will constitute individually novel contributions to statistics, ML, and HPC computer science, contributing new results to climate and space weather science that would not otherwise have been possible.

Risk Assessment and Mitigation

Part of the project mitigation will be to decouple deliverables: much of the inference, including distributed inference, can be developed and tested offline before coupling into HPC codes. The ML component may be pursued independently without necessarily being implemented in-situ.

Scalability. Potential barriers to scalability are whether passing summary statistics will incur too much communication overhead, or VI over a large number of SGP pseudo-data parameters will slow down the simulation unacceptably. The initial mitigation plan is to reduce the number of pseudo-data until performance is acceptable. This will reduce fidelity in the statistical model (controllably, with an automatic probabilistic estimate of the approximation error introduced), which may be acceptable in some applications with less complex structure. ML-guided summary statistics are one way we hope to circumvent potential communications barriers, by reducing distributed data exchange to the minimum needed to preserve validity of the inference.

Spatiotemporal model complexity. Our statistical model will take effort to get running in-situ in a distributed, streaming HPC setting. If this takes more time than anticipated, we will take a two-pronged-approach: develop the full model on (smaller) offline data, and reduce the complexity of the in-situ model by focusing on just the spatial part, neglecting nonstationary changes in time; we can construct steady-state simulations to produce data lacking strong time trends.

In-situ software framework and HPC coupling. A potential risk is the relative youth of the Julia language.[9] Julia has exhibited significant maturity for HPC: Petascale Bayesian VI on DOE leadership computing;[76] competitive ML performance with Python frameworks;[45,31] and selected for an E3SM-class climate model with beyond-Fortran performance.[69,78] But it has not yet been coupled to a legacy DOE HPC code. We have compiler experts to mitigate risk, but in case of unanticipated challenges that could delay the project, we will implement inference using a more orthodox approach: Python (Ascent),[59] with C++ ports of code hotspots. This will not achieve the full performance and generality possible with Julia, but provides a path to in-situ inference. A rapid technical risk assessment at the project start will inform this decision.

Project plan

Urban will lead the project. **Lawrence** will lead the inference with **Grosskopf** on spatiotemporal and variational methods, and **Dorn** on distributed inference with summary statistics. **Biswas** will

lead the in-situ framework with **McCormick**, **Stelle**, and **Dutta** on high-level programming language abstractions, parallelism, and in-situ expertise. **Oyen**, **Urban**, and **Dorn** on ML-guided inference. **Urban** and **Van Roekel** will lead the climate application with **Wolfe** on E3SM coupling; **Jordanova**, **Henderson**, and **Morley** will lead the space weather application.

The project will have three R&D areas that will converge as the project progresses. These are (1) using the probabilistic model to study science questions, (2) scalable HPC-ready inference algorithms, and (3) the in-situ computational framework and high-level programming abstraction layer.

Year 1. Begin developing the inference algorithm for the full statistical model on offline simulation data, without requiring scalability. Initial focus is on fitting the the spatial part of the model, neglecting time. Start developing a distributed (but not streaming) SGP model. Begin coupling the Julia runtime environment into one of the target HPC codes, and implement within it a simple inference algorithm, such as linear regression or fitting (spatially-independent, stationary) extreme value distributions. We will begin developing shared memory concurrency abstractions.

Year 2. Improve the main scientific analysis for both applications by including time variations into the probabilistic model. Begin adding streaming inference to our distributed offline algorithm and improve its scalability with linear-complexity and/or Krylov methods. Begin developing ML summary statistics through offline synthetic training. Insert a simplified version of a distributed GP model in-situ in a model as an intermediate deliverable. Continue developing concurrency.

Year 3. Improve the scientific analysis by including covariate information (large-scale state data). We will complete the scaling of the distributed, streaming SGP model and insert it in-situ. We will complete the ML summary statistics model (offline). We will extend the high-level abstraction layer to remove some of the initial hard-coding of algorithms (e.g. parallel linear algebra).

Data management plan

We will request 250 TB of campaign storage on LANL IC for our new simulations, statistical models, and training data (existing model output), in standard formats such as NetCDF. Code will be open-sourced on Github and trained statistical models made available on Zenodo.

Transition plan

G. Shipman is LANL's point of contact (POC) for computational partnerships between applied math, computer science, and domain science in the ASCR program, a natural follow-on to this DR. P. McCormick (a team member) is LANL's ASCR POC for computer science and can represent the proposal there. In-situ analysis was the subject of a recent ASCR workshop,[46] so this project will be well-positioned for follow-on funding. R. Friedel (CSES Director) will help connect to potential opportunities in NASA and other agencies interested in the space weather problem. This effort also has strong connection to DOE's programs in climate and earth sciences (BER; E. Hunke), which is prioritizing increasing prediction fidelity to assess climate impacts. S. Vander Wiel serves as the Advanced Certification Campaign UQ Project Lead, and will help identify applications in the weapons program will be suitable for follow-on work.

Budget Request

$1.615M/y will fund 13 TSMs (at 10-30% levels) and 3.5 postdocs. We request travel (~$30K/y) and M&S (~$30K/y); M&S covers the purchase of computers for postdocs and publication costs.

Glossary of acronyms

ASCR = Advanced Scientific Computing Research (DOE program)

BER = Biological and Environmental Research (DOE program)

BBF = bursty bulk flow (space weather event)

DNN = deep neural network

CSES = Center for Space and Earth Sciences (LANL institute)

E3SM = Energy Exascale Earth System Model (DOE climate model)

ELBO = evidence lower bound (variational objective function to maximize)

ECP = Exascale Computing Project (DOE program)

GEV = generalized extreme value (distribution)

GP = Gaussian process (spatiotemporal statistical model)

GPU = graphics processing unit (hardware accelerator)

HPC = high performance computing

IC = Institutional Computing (LANL)

LANL = Los Alamos National Laboratory

M&S = materials and supplies

MHD = magnetohydrodynamics

ML = machine learning

NN = neural network

MPI = Message Passing Interface (HPC distributed communication protocol)

POC = point of contact

PDF = probability density function

SGP = sparse Gaussian process

SHIELDS = Space Hazards Induced near Earth by Large Dynamic Storms (space weather model)

SSGP = streaming sparse Gaussian process

TSM = Technical Staff Member

UQ = uncertainty quantification

VI = variational inference

Citations

[1] J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, and M. Petersen, "An image-based approach to extreme scale in situ visualization and analysis", *International Conference for High Performance Computing, Networking, Storage and Analysis* (2014).
[2] E. Banijamali, A.-H. Karimi, and A. Ghodsi, "Deep variational sufficient dimensionality reduction", arXiv:cs/1812.07641 (2018).
[3] S. Banerjee, B.P. Carlin, and A.E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC (2014).
[4] A.C. Bauer, B. Geveci, and W. Schroeder: "The ParaView Catalyst User's Guide v2.0". Kitware, Inc. (2015).
[5] M. Bauer, M. van der Wilk, and C.E. Rasmussen, "Understanding probabilistic sparse Gaussian process approximations", *Neural Information Processing Systems* (2016).
[6] T. Besard, C. Foket, and B. De Sutter, "Effective extensible programming: Unleashing Julia on GPUs", arXiv:cs/1712.03112 (2017).
[7] J. Berckmans, T. Woollings, M.-E. Demory, P.-L. Vidale, and M. Roberts, "Atmospheric blocking in a high resolution climate model: Influences of mean state, orography and eddy forcing", *Atmospheric Science Letters* **14**, 34 (2013).
[8] J. Bezanson, J. Chen, S. Karpinski, V. Shah, A. Edelman, "Array operators using multiple dispatch: A design methodology for array implementations in dynamic languages", *ARRAY '14* (2014).
[9] J. Bezanson, A. Edelman, S. Karpinski, and V.B. Shah, "Julia: A fresh approach to numerical computing", *SIAM Review* **59**, 65 (2017).
[10] A. Biswas, S. Dutta, J. Pulido, and J. Ahrens, "In-situ data-driven adaptive sampling for large-scale simulation data summarization", *In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization* (2018)
[11] A. Biswas, S. Dutta, J.M. Patchett, E.C. Lawrence, and J.P. Ahrens, "Probabilistic data-driven sampling via multi-riteria importance analysis", *IEEE Visualization* (2019), submitted (LA-UR-19-23056).
[12] D.M. Blei, A. Kucukelbir, and J.D. McAuliffe, "Variational inference: A review for statisticians", *Journal of the American Statistical Association* **112**, 859 (2017).
[13] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond Euclidean data", *IEEE Signal Processing Magazine* **34**, 18 (2016).
[14] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics", *International Conference on Learning Representations* (2016).
[15] T.D. Bui, C. Nguyen, and R.E. Turner, "Streaming sparse Gaussian process approximations", *Neural Information Processing Systems* (2017).
[16] P. de Castro and T. Dorigo, "INFERNO: Inference-aware neural optimisation", arXiv:stat/1806.04743 (2018).
[17] S. Chen, P.B. Gibbons, M. Kozuch, V. Liaskovitis, G.E. Blelloch, B. Falsafi, L. Fix, N. Hardavellas, T.C. Mowry, and C. Wilkerson, "Scheduling threads for constructive cache sharing on CMPs", *Proceedings of the Nineteenth Annual ACM Symposium on Parallel Algorithms and Architectures* (2007).
[18] C.-A. Cheng and B. Boots, "Variational inference for Gaussian processes with linear complexity", *Neural Information Processing Systems* (2017).
[19] A. Choudhury, P.B. Nair, and A.J. Keane, "A data parallel approach for large-scale Gaussian process modeling", *SIAM International Conference on Data Mining* (2002).
[20] J. Cohen, J.A. Screen, J.C. Furtado, M. Barlow, D. Wittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, and J. Jones, "Recent Arctic amplification and extreme mid-latitude weather", *Nature Geoscience* **7**, 627 (2014).
[21] D. Cooley, D. Nychka, and P. Naveau, "Bayesian modeling of extreme precipitation return levels", *Journal of the American Statistical Association* **102**, 824 (2012).
[22] J. Cornelis, S. Cools, and W. Vanroose, "The communication-hiding conjugate gradient method with deep pipelines", arXiv:cs/1801.04728 (2019).
[23] A.C. Damianou and N.D. Lawrence, "Deep Gaussian processes", *International Conference on Artificial Intelligence and Statistics* (2013).
[24] A.M. DeGennaro, N.M. Urban, B.T. Nadiga, and T. Haut, "Model structural inference using local dynamic operators", *International Journal of Uncertainty Quantification* **9**, 59 (2019).
[25] M.P. Deisenroth and J.W. Ng, "Distributed Gaussian processes", *International Conference on Machine Learning* (2015).
[26] S. Dutta, C.M. Chen, G. Heinlein, H.W. Shen, and J.P. Chen. "In situ distribution guided analysis and visualization of transonic jet engine simulations", *IEEE Transactions on Visualization and Computer Graphics*, **23**, 811 (2017).

[27] S. Dutta, H.W. Shen, and J.P. Chen. "In situ prediction driven feature analysis in jet engine simulations", *IEEE Pacific Visualization Symposium* (2018).
[28] S. Dutta, J. Woodring, H.W. Shen, J.P. Chen, and J. Ahrens, "Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets", *IEEE Pacific Visualization Symposium* (2017).
[29] P.R. Eller and W. Gropp, "Scalable non-blocking preconditioned conjugate gradient methods", *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (2016).
[30] N. Fabian, K. Moreland, D. Thompson, A.C. Bauer, P. Marion, B. Geveci, M. Rasquin, and K.E. Jansen: "The Paraview Coprocessing Library: A scalable, general purpose in situ visualization library", *IEEE Symposium on Large-Scale Data Analysis and Visualization* (2011).
[31] K. Fischer and E. Saba, "Automatic full compilation of Julia programs and ML models to cloud TPUs", *Neural Information Processing Systems* (2018).
[32] Flang Compiler, https://github.com/flang-compiler
[33] Y. Gal, Y. Chen, and Z. Gharamani, "Latent Gaussian processes for distribution estimation of multivariate categorical data", *International Conference on Machine Learning* (2015).
[34] Y. Gal, M. van der Wilk, and C.E. Rasmussen, "Distributed variational inference in sparse Gaussian process regression and latent variable models", *Neural Information Processing Systems* (2014).
[35] R.B. Gramacy, J. Niemi, and R.M. Weiss, "Massively parallel approximate Gaussian process regression", *SIAM/ASA Journal of Uncertainty Quantification* **2**, 564 (2014).
[36] R. Gramacy, D. Bingham, J.P. Holloway, M. Grosskopf, C.C. Kuranz, E. Rutter, M. Trantham, and R.P. Drake, "Calibrating a large computer experiment simulating radiative shock hydrodynamics", *Annals of Applied Statistics* **9**, 1141 (2015).
[37] A. Grigorievskiy, N. Lawrence, and S. Särkkä, "Parallelizable sparse inverse formulation Gaussian processes (SpInGP), *International Workshop on Machine Learning for Signal Processing* (2017).
[38] M. Grosskopf, D. Bingham, M. Adams, W.D. Hawkins, and D. Perez-Nunez, "Generalized computer model calibration for radiation transport simulation", *Technometrics*, in revision (2019).
[39] R. Guhaniyogi et al., "Distributed kriging: A divide-and-conquer Bayesian approach to large-scale kriging", LANL In-Situ Inference seminar, April 2, 2019.
[40] J. Guiness and D. Hammerling, "Compression and conditional emulation of climate model output", *Journal of the American Statistical Association* **113**, 56 (2018).
[41] K. Heitmann, S. Habib, H. Finkel, N. Frontiere, A. Pope, V. Morozov, S. Rangel, E. Kovacs, J. Kwan, N. Li, S. Rizzi, J. Insley, V. Vishwanath, T. Peterka, D. Daniel, P. Fasel, and G. Zagaris, "Large-scale simulations of sky surveys", *Computing in Science & Engineering* **16**, 14 (2014).
[42] T.N. Hoang, Q.M. Hoang, B.K. Hsiang Low, "A distributed variational inference framework for unifying parallel sparse Gaussian process regression models", *International Conference on Machine Learning* (2016).
[43] T.N. Hoang, Q.M. Hoang, K. Hsiang Low, and J. How, "Collective online learning of Gaussian processes in massive multi-agent systems", arXiv:cs/1805.09266 (2018).
[44] M. Innes, "Don't unroll adjoint: Differentiating SSA-form programs", *Neural Information Processing Systems* (2018).
[45] M. Innes, E. Saba, K. Fischer, D. Gandhi, M.C. Rudilosso, N.M. Joy, T. Karmali, A. Pal, and V. Shah, "Fashionable modelling with Flux", *Neural Information Processing Systems* (2018).
[46] DOE In-Situ Data Workshop, January 28–29, 2019, https://www.orau.gov/insitudata2019/
[47] J. Jeong, S. Castruccio, P. Crippa, and M.G. Genton, "Reducing storage of global wind ensembles with stochastic generators", *Annals of Applied Statistics* **12**, 490 (2018).
[48] B. Jiang, T.-Y. Yu, C. Zheng, and W.H. Wong, "Learning summary statistics for approximate Bayesian computation via deep neural network", *Statistica Sinica* **27**, 1595 (2017).
[49] T. Jiang, K. Evans, M. Branstetter, P. Caldwell, R. Neale, P.J. Rasch, Q. Tang, and S. Xie, "Northern hemisphere blocking in ~25-km-resolution E3SM v0.3 atmosphere-land simulations", *Journal of Geophysical Research—Atmospheres* **124**, 2465 (2019).
[50] V.K. Jordanova, G.L. Delzanno, M.G. Henderson, H.C. Godinez, C.A. Jeffery, E.C. Lawrence, S.K. Morley, J.D. Moulton, L.J. Vernon, J.R. Woodroffe, T.V. Brito, M.A. Engel, C.S. Meierbachtol, D. Svyatsky, Y. Yu, G. Tóth, D.T. Welling, Y. Chen, J. Haiducek, S. Markidis, J.M. Albert, J. Birn, M.H. Denton, and R.B. Horne, "Specification of the near-Earth space environment with SHIELDS", *Journal of Atmospheric and Solar-Terrestrial Physics* **177**, 148 (2018).
[51] V.K. Jordanova, S. Zaharia, and D.T. Welling, "Comparative study of ring current development using empirical, dipolar, and self-consistent magnetic field simulations", *Journal of Geophysical Research* **115**, A00J11 (2010).
[52] Julia: Parallel Computing in Julia, https://docs.julialang.org/en/v1/manual/parallel-computing/index.html

[53] B.-M. Kim, S.-W. Son, S.-K. Min, J.-H. Jeong, S.-J. Kim, X. Zhang, T.-t. Shim, and J.-H. Yoon, "Weakening of the stratospheric polar vortex by Arctic sea-ice loss", *Nature Communications* **5**, 4646 (2014).
[54] Kitsune Compiler, https://github.com/lanl/kitsune
[55] N. Klein, "Emulating large simulation outputs for visualization", LA-UR-19-20948 (2019).
[56] M. Kretschmer, D. Coumou, L. Agel, M. Barlow, E. Tziperman, and J. Cohen, "More-persistent weak stratospheric polar vortex states linked to cold extremes", *Bulletin of the American Meterological Society* **99**, 49 (2018).
[57] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D.M. Blei, "Automatic differentiation variational inference", *Journal of Machine Learning Research* **18**, 1 (2017).
[58] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, Prahbat, and M. Houston, "Exascale deep learning for climate analytics", *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis* (2018).
[59] M. Larsen, J. Ahrens, U. Ayachit, E. Brugger, H. Childs, B. Geveci, and C. Harrison, "The ALPINE In Situ Infrastructure: Ascending from the Ashes of Strawman", *In Situ Infrastructures on Enabling Extreme-Scale Analysis and Visualization* (2017).
[60] M. Larsen, A. Woods, N. Marsaglia, A. Biswas, S. Dutta, C. Harrison, and H. Childs, "A flexible system for in situ triggers", *In Situ Infrastructures for Enabling Extreme-scale Analysis and Visualization* (2018).
[61] E. Lawrence, K. Heitmann, J. Kwan, A. Upadhye, D. Bingham, S. Habib, D. Higdon, A. Pope, H. Finkel, and N. Frontiere, "The Mira-Titan Universe. II. Matter power spectrum emulation", *Astrophysical Journal* **847**, 1 (2017).
[62] F. Li, M. Villani, and R. Kohn, "Flexible modeling of conditional distributions using smooth mixtures of asymmetric student *t* densities", *Journal of Statistical Planning and Inference* **140**, 3638 (2010).
[63] J. Lin, M. Wen, D. Meng, X. Liu, A. Nukada, and S. Matsuoka, "Optimizing preconditioned conjugate gradient on TaihuLight for OpenFOAM", *Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (2019).
[64] Y. Lin, S. Han, H. Mao, Y. Wang, and W. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training", *International Conference on Machine Learning* (2018).
[65] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gaussian process meets big data: A review of scalable GPs", arXiv:stat:1807/010165 (2019).
[66] L. Liu and L. Liu, "Amortized variational inference with graph convolutional networks for Gaussian processes", *Proceedings of Machine Learning Research* (2019).
[67] Q. Liu, J. Logan, Y. Tian, H. Abbasi, N. Podhorszki, J.Y. Choi, S. Klasky, R. Tchoua, J. Lofstead, R. Oldfield, M. Parashar, N. Samatova, K. Schwan, A. Shoshani, M. Wolf, K. Wu, and W. Yu, "Hello ADIOS: The challenges and lessons of developing leadership class I/O frameworks", *Concurrency and Computation: Practice and Experience* **26**, 1453 (2014).
[68] K. Myers, E. Lawrence, M. Fugate, C.M.K. Bowen, L. Ticknor, J. Woodring, J. Wendelberger, and J. Ahrens, "Partitioning a large simulation as it runs", *Technometrics* **58**, 329 (2016).
[69] J. Neasbitt, "NPS researchers partner on next generation climate model", Naval Postgraduate School press release, March 4, 2019, https://my.nps.edu/-/nps-researchers-partner-on-next-generation-climate-model
[70] D. Peel and G.J. McLachlan, "Robust mixture modelling using the *t* distribution", *Statistics and Computing* **10**, 339 (2000).
[71] H. Peng, S. Zhe, and Y. Qi, "Asynchronous distributed variational Gaussian processes for regression", arXiv:stat/1704.06735 (2017).
[72] T.H. Pham, J.T. Ormerod, and M.P. Wand, "Mean field variational Bayesian inference for nonparametric regression with measurement error", *Computational Statistics & Data Analysis* **68**, 375 (2013).
[73] C. Rackauckas, "Why Numba and Cython are not substitutes for Julia" (2018), http://www.stochasticlifestyle.com/why-numba-and-cython-are-not-substitutes-for-julia/
[74] C. Rackauckas and Q. Nie, "Confederated modular differential equation APIs for accelerated algorithm development and benchmarking", arXiv:cs/1807.06430 (2018).
[75] C. Rackauckas, M. Innes, Y. Ma, J. Bettencourt, L. White, and V. Dixit, "DiffEqFlux.jl — A Julia library for neural differential equations" (2019), https://julialang.org/blog/2019/01/fluxdiffeq
[76] J. Regier, K. Pamnany, K. Fischer, A. Noack, M. Lam, J. Revels, S. Howard, R. Giordano, D. Schlegel, J. McAuliffe, R. Thomas, and Prabhat, "Cataloging the visible universe through Bayesian inference at petascale", *Journal of Parallel and Distributed Computing* **127**, 89 (2019).
[77] J. Revels, T. Besard, V. Churavy, B. De Sutter, and J.P. Vielma, "Dynamic automatic differentiation of GPU broadcast kernels", arXiv:cs/1810.08297 (2018).
[78] T. Schneider, "Earth System Modeling 2.0: Toward accurate and actionable climate predictions with quantified uncertainties", LANL In-Situ Inference seminar, March 6, 2019.

[79] J.A. Screen, I. Simmonds, C. Deser, and R. Tomas, "The atmospheric response to three decades of observed Arctic sea ice loss", *Journal of Climate* **26**, 1230 (2013).
[80] T.B. Schardl, W.S. Moses, and C.E. Leiserson., "Tapir: Embedding Fork-Join Parallelism into LLVM's Intermediate Representation", *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (2017).
[81] G. Stelle, W.S. Moses, S.L. Olivier, and P. McCormick. "OpenMPIR: Implementing OpenMP Tasks with Tapir", *LLVM Compiler Infrastructure in HPC* (2017).
[82] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs", *Neural Information Processing Systems* (2005).
[83] E. Snelson and Z. Gharamani, "Local and global sparse Gaussian process approximations", *International Conference on Artificial Intelligence and Statistics* (2007).
[84] B.V. Srinivasan, Q. Hu, N.A. Gumerov, R. Murtugudde, and R. Duraiswami, "Preconditioned Krylov solvers for kernel regression", *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2014).
[85] M.K. Titsias, "Variational learning of inducing variables in sparse Gaussian processes", *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics* (2009).
[86] G. T'oth, B. van der Holst, I.V. Sokolov, D.L.D. Zeeuw, T.I. Gombosi, et al., "Adaptive numerical algorithms in space weather modeling", *Journal of Computational Physics* **231**, 870 (2012).
[87] A. Verdin, B. Rajagopalan, W. Kleiber, G. Podestá, and F. Bert, "BayGEN: A Bayesian space-time stochastic weather generator", *Water Resources Research*, in press (2019), DOI: 10.1029/2017WR022473
[88] V. Vishwanath, M. Hereld, V. Morozov, and M.E. Papka, "Topology-aware data movement and staging for I/O acceleration on Blue Gene/P supercomputing systems", *International Conference for High Performance Computing, Networking, Storage and Analysis* (2011).
[89] H. von Storch and F.W. Zwiers, *Statistical Analysis in Climate Research* (1999).
[90] B. Whitlock, J.M. Favre, and J.S. Meredith, "Parallel in situ coupling of simulation with a fully featured visualization system", *11th Eurographics conference on Parallel Graphics and Visualization* (2011).
[91] J. Woodring, M. Petersen, A. Schmeißer, J. Patchett, J. Ahrens, and H. Hagen, "In situ eddy analysis in a high-resolution ocean climate model", *IEEE Transactions on Visualization and Computer Graphics* **22**, 857 (2016).
[92] A. Woodward, "The US is suffering through a polar vortex. Paradoxically, we may have global warming to thank for that", *Business Insider*, January 31, 2019.
[93] Z.J. Zhang and K. Duraisamy, "Machine learning methods for data-deriven turbulence modeling", *AIAA Computational Fluid Dynamics Conference* (2015).

Appendices

Computing Resource Needs

Although the motivation for the project is exascale computing, most of the development will be on less-expensive model configurations, since it is too demanding to routinely run the highest resolution configurations every time we modify the in-situ software or statistical model.

The E3SM low-resolution configuration (~3.6 million atmospheric degrees of freedom per variable) requires ~23,000 core-hours per simulated year. A simulation sufficient to capture the statistics of extreme events would be ~100 simulated years, or 2.3 million core-hours. However, a simulation of this length is only needed for the science campaigns, which only need to be performed perhaps twice over the duration of the project. Most of the algorithm development can be tested on much shorter (say, 5-year) integrations. We will budget for 300 simulated years at low resolution, or 6.9 million core-hours over the span of the 3-year project. The E3SM high-resolution configuration (~58 million atmospheric degrees of freedom) requires ~2.8 million core-hours per simulated year. We will only demonstrate inference on 5 simulated years at high resolution, requesting an 14 million core-hours from ASCR Leadership Computing Challenge (ALCC).

SHIELDS requires 11,000 core hours for 1 simulated day, which is the typical duration of a geomagnetic storm. To resolve bursty bulk flows the SHIELDS framework will require a minimum spatial resolution of 1/8 Earth radii in the near-tail region. This can be achieved using a targeted simulation grid of approximately 4 million cells in the global MHD component of SHIELDS. This configuration requires ~11,000 core-hours per simulated day. For the final science campaign we will use higher resolution simulations (~6M cells, ~17k core-hours/day). To capture the statistics of extreme BBFs we will require sets of simulations for different solar wind drivers, dipolar tilt angles, and preconditioning levels. Constructing upstream boundary conditions (solar wind drivers) to capture different types of preconditioning, for a range of dipole tilt angles, will give a set of initial model configurations that can be used for subsequent simulations of the effects of solar wind types known to produce extreme activity, such as interplanetary coronal mass ejections. We here budget for a total of 45 stormtime simulations (135 days) at 6M cell resolution, corresponding to ~2.3 million core-hours for the science campaign. Algorithm development can be tested on individual storm simulations (~3 days, or 33k core-hours each) using the 4M cell configuration. We here budget for 40 stormtime simulations at 4M cell resolution for algorithm development and testing, corresponding to ~1.3 million core-hours over the span of the project. The total computing resources required for the space weather modeling is 3.6 million core-hours.

People

In addition to the technical capability described in the narrative, this project will provide a tremendous opportunity to develop the staff and collaborations that are necessary for the Lab's future success. All members of the project will be able to advance as scientists in the pursuit of this project's goals. Further, LANL will have the opportunity to establish a leadership position at the nexus of high performance computing and data analysis.

This project will provide support for three new postdocs and one existing postdoc. By the nature of the project, this support will provide an opportunity to recruit new potential new staff members with multidisciplinary skills critical to Lab success. One current (Dutta) and one new postdoc will support the computer science mission of the proposal. These postdocs will develop

skills implementing advanced statistical analyses in-situ. This will provide experience with statistics and machine learning outside the typical CS research framework. Two other postdocs will support the statistics and machine learning missions. These postdocs will gain experience developing new methods with the constraints and advantages provided by modern high performance computing. Both sets of post-docs will have a set of skills vital to the future of the Laboratory.

The proposal will also provide similar support for numerous early career staff members. This staff includes Biswas, Dorn, Grosskopf, and Stelle. This support will promote cross-disciplinary skills. The statistics staff (Dorn, and Grosskopf) will develop the computational skills and experience needed to support advanced data analysis in future Lab computing environments. The computer science staff (Biswas and Stelle) will gain experience implementing and using advanced data analysis techniques and creating the support for these methods. Collaborations between these early career staff have the potential to bear fruit for the Lab for many years. Early career staff will also be placed in positions of leadership where possible. They will be given opportunities to represent the project internally, as part of project reviews, and externally at conferences. They will be encouraged to take the lead on developing follow-on and spin-off projects.

Senior leadership will also have the opportunity for career growth. Urban and Lawrence are both leaders in uncertainty quantification. LANL is also at the forefront of this field. The project will strengthen that position nationally and internationally by advancing uncertainty quantification for the exascale era. McCormick is a leader in programming models and parallel computing. This project will provide an opportunity for to advance his expertise in a new direction.

This proposal brings together a cross-disciplinary team from CCS, T, and ISR divisions. Within CCS, the proposal will strengthen a number of existing collaborations. PI Urban has worked extensively with CCS-6 in numerous areas and with the climate team from T-3. Likewise, Co-PI Lawrence has worked extensively with CCS-7 and the space weather modeling team from ISR-1. With this foundation, it should be easy to develop the total collaboration required for the project. These collaborations will strengthen both science and data analysis capabilities at the LANL.

Finally, Urban and Lawrence have already begun to engage the larger academic community on this problem. We are currently running a seminar series sponsored by ISTI on the topic of in-situ inference. The series hosts experts working on scalable machine learning methods, including Gaussian process modeling, and fast inference and estimation methods. Although not a set of formal collaborations, this series has connected us with researchers at the cutting edge of relevant methodology and has begun to lay the groundwork for future collaboration with outside partners. This will help establish LANL as a leader in the important field, and could be continued or expanded into a recurring LANL-led conference.

Biosketches, Level of Effort, and Project Roles

Nathan Urban (PI, CCS-2, ~0.3 FTE) has 12 years of experience in statistical uncertainty quantification (UQ), reduced-order modeling, climate science, and climate impacts analysis. His activities at the lab connect physical science, numerical simulation, Bayesian UQ, and decision analysis to practical problems of national and policy interest. His research themes include multimodel or model “structural” uncertainties, reduced order modeling for UQ, machine learning to improve numerical models, uncertainty analysis of computationally expensive numerical models, climate feedbacks and sea level rise, climate adaptation and risk management, and infrastructure

network analysis and resiliency design optimization. As a DOE Office of Science Early Career Researcher, he has played leadership roles both in LANL internal strategy for climate impacts, sea level and coastal science, climate resilience, and national security, as well as DOE Office of Science program development activities in the BER and ASCR divisions. He will lead the entire project, contributing to both the inference and software sides of the project.

Earl Lawrence (Co-PI, CCS-6, ~0.25 FTE) has extensive experience in uncertainty quantification and simulation-based inference since joining the Lab in 2005 and has published numerous papers in this area. He is an expert in the application of Bayesian inference and spatiotemporal models for inference using simulations. Applications include nuclear weapons, cosmology, space weather, and power grids. He led a successful LDRD on uncertainty quantification for power grids and led the UQ efforts for several LDRD projects. He is currently the uncertainty quantification lead on two Office of Science projects on cosmology and nuclear theory. He is on the editorial board of the SIAM/ASA Journal on Uncertainty Quantification and is the Chair of the American Statistical Association's Interest Group on Uncertainty Quantification. He will lead the inferential development of the scalable Gaussian process and extreme value modeling.

Ayan Biswas (Co-PI, CCS-7, ~0.3 FTE) is a data scientist with a Ph.D. in Computer Graphics and Data Visualization. His Ph.D. work focused on large-scale high-dimensional and multivariate datasets and uncertainty visualization. He has extensive experience in information theory, HPC and data modeling. He is currently the sampling lead for the ALPINE project under Exascale Computing Project (ECP) that focuses on massive scale parallel distributed in-situ sampling algorithms. His knowledge and expertise in such related topics allow him to understand the gaps in the existing literature and help fill that through this cutting-edge research. He will lead the development of the computational foundations for the in-situ framework.

Diane Oyen (Co-I, CCS-3, ~0.15 FTE) develops machine learning algorithms for pattern discovery in scientific data. She applies a variety of ML algorithms — including deep neural networks, autoencoders, and probabilistic graphical models — to diverse data sets, including spatio-temporal physics simulations, image analysis, spectroscopy, and cybersecurity. Oyen has extensive research in simultaneous learning of multiple high-likelihood machine learning models through sharing of summary statistics among models for efficient training. Oyen will contribute to the deep sparse Gaussian process with summary statistics passed among local inference algorithms.

Mary Frances Dorn (Co-I, CCS-6, ~0.25 FTE) has experience modeling spatiotemporal data and developing and applying machine learning algorithms in a variety of applications. As part of her dissertation research, she developed classification methods that, instead of learning an unknown underlying correlation structure, computes simple low-dimensional summaries of the complex data that allowed for scaling to high-dimensional problems. Since joining the Lab in 2017, she has worked on modeling spatiotemporal data with applications including power outage forecasting during extreme weather events and employee injury incidents around the lab, and uncertainty quantification for machine learning algorithms applied to nuclear non-proliferation detection and physics-based models of material characteristics. She will work on methods for the machine-learning guided inference.

Michael Grosskopf (Co-I, CCS-6, ~0.3 FTE) is an early career staff member with expertise in uncertainty quantification and inference with expensive simulators including emulations of models with multivariate outcomes and modeling latent functions using Gaussian process regression with non-Gaussian measurement distributions. Additionally, he is experienced in large-scale distributed computing with simulation models and working with domain scientists in computational

physics applications. He has also worked in applications with online data collection and inference for anomaly detection.

Soumya Dutta (Co-I, CCS-7, ~0.5 FTE) is a postdoc research associate in the Data Science at Scale group under CCS-7 at Los Alamos National Lab. He has a Ph.D. in computer science and engineering and has expertise in multivariate data analysis, in situ analysis and visualization, and statistical data analytics which makes him suitable for this project. He has published in situ analysis papers in important visualization conferences and after joining LANL, he continues to contribute heavily towards the in situ efforts of Data Science at Scale (CCS-7) group. He will contribute to the development of the computational foundations of the in situ framework.

George Stelle (Co-I, CCS-7, ~0.25 FTE) is an expert in the design and implementation of programming languages and compilers. By representing concurrency semantics in the compiler, George will enable analyses and optimizations that would otherwise be missed. Applying George's skill set to the concurrency primitives in Julia, we expect performance and portability for our in-situ codes not available elsewhere. Having an expert in programming languages and compilers will also help mitigate the risk of using a relatively new language.

Pat McCormick (Co-I, CCS-7, ~0.1 FTE) is a senior computer scientist at Los Alamos National Laboratory (LANL) with over 25 years of experience in high-performance computing, and is well known and respected for his work in programming models, early GPU programming, data visualization, and parallel systems. At LANL, he serves as the Programming Models team leader in CCS-7, and as the Program Manager for LANL Office of Science (SPO-SC), Advanced Scientific Computing Research. McCormick has authored or coauthored numerous papers and has extensive project management experience including serving as the Deputy Director for the Software Technology area of the Exascale Computing Project from 2015-2017. He will be providing technical guidance on the research and development aspects of the in situ framework.

Luke Van Roekel (Co-I, T-3, ~0.25 FTE) has extensive experience simulating the climate across scales, from the subgrid (<1km) to planetary scale. He is a co-lead developer of the LANL developed Model for Prediction Across Scales-Ocean and a science group co-lead for the Energy Exascale Earth System Model (E3SM) project, which is DOEs new Earth System Model that was built to address DOE mission relevant questions. On this project, he will lead the domain science application for climate, helping design simulations, and setup and configure the model. He will also advise CCS staff on how to interface the in-situ analysis code with E3SM, and will provide output for the first year to conduct the analysis offline to verify the newly developed statistical algorithms can interface with the E3SM output.

Jonathan Wolfe (Co-I, T-3, ~0.25 FTE) has extensive experience in parallel high-performance software infrastructure and integration for climate change and energy related numerical simulation, including research, design and development, verification and validation, production, customer implementation. He has scientific model development/applications for industrial and national needs, including geothermal reservoirs models, flow through heterogeneous media, nuclear weapons stockpile stewardship, safety and waste-processing reactors, and wind forecasting. He will contribute to the analysis and interpretation of climate simulations and data.

Michael Henderson (Co-I, ISR-1, ~0.1 FTE) has worked at Los Alamos National Laboratory since 1994 as a space physicist working on problems associated with magnetospheric storms, substorms and space weather. He has worked with data from a wide variety of space-borne instrumentation including: Viking/UVI, POLAR/CEPPAD, POLAR/CAMMICE,

CLUSTER/RAPID, LANL Geosynchronous CPA, SOPA, MPA and ESP instruments, IMAGE/MENA, Van Allen Storm Probes Mission, Magnetospheric Multiscale Mission, and mostly recently in planning activities for the CONNEX magnetospheric mapping mission. Dr. Henderson is an expert in the analysis of global auroral imagery and geosynchronous energetic particle and plasma data and has extensive experience working with many other ground-based and space-based datasets. His research interests currently focus on: Assessing impacts of extreme space weather events (Carrington-Class Geomagenetic Storms) on power grid infrastructure; The relationship between storms, substorms, Steady Magnetospheric Convection events (SMCs), and sawtooth events; Simulation of inner magnetospheric energetic particle dispersion patterns; The extraction of radial diffusion transport coefficients from Van Allen Probes and LANL/GEO energetic particle datasets using advanced data assimilation techniques. He is the author of the LanlGeoMag library which provides extensive routines for very precise coordinate transformations, orbit propagation, magnetic field line tracing, and adiabatic invariant calculations. LanlGeoMag is a core library used extensively in the DREAM space weather model (radiation belt data assimilation), Van Allen Probes, Magnetospheric Multiscale Mission and other projects at LANL. Dr. Henderson is currently leading the large LDRD/DR space weather modeling effort (in its last year): "Impacts of Extreme Space Weather Events on Power Grid Infrastructure: Physics-Based Modeling of Carrington-Class Geomagnetic Storm Events" which utlizes the same space weather modeling infrastructure (Space Weather Modeling Framework or SWMF) we propose to use here. In the current proposal, Dr. Henderson will mainly be involved in the development, analysis and interpretation of the magnetospheric simulations in the space weather portion of the work and in the integration of the in-situ algorithms as modules for use in the SWMF.

Vania Jordanova (Co-I, ISR-1, ~0.1 FTE) has over twenty years of experience in theoretical, observational, and numerical studies of the Earth's magnetosphere, ring current and radiation belt dynamics, wave-particle interactions, and processes that couple the ionospheric and magnetospheric regions. She created a state-of-the-art kinetic ring current-atmosphere interactions model (RAM) that simulates the transport of hot (kiloelectron-volt) ions and electrons in realistic electric and magnetic fields, taking into account all key source and loss processes. The RAM code is one of the main modules of the space weather modeling framework which aims at specifying Space Hazards Induced near Earth by Large Dynamic Storms (SHIELDS). She has substantial leadership experience as the PI on many NASA and NSF projects and the LDRD-DR SHIELDS project, a 2017 R&D 100 Award winner. She has more than 130 scientific publications and has presented more than 40 invited talks at international conferences. She served on many NASA, NSF and DOE review panels, the NSF/Geospace Environment Modeling (GEM) Program Steering Committee, and was a member of the NASA Magnetosphere Working Group on Advanced Computational Exploration (ACCEHS, 2010). She will assist with the SHIELDS space weather simulations and their interpretation. She will also contribute to the analysis and interpretation of magnetospheric plasma data and model/data validation. She will assist with the application of the results from this effort to existing and future internal and external programs.

Steve Morley (Co-I, ISR-1, ~0.3 FTE) has expertise in the large-scale responses of the magnetosphere and ionosphere to changes in the solar wind. His work has included development of novel models of ionospheric convection and substorm occurrence, and validation against observations, and development of inner magnetosphere and radiation belt simulations. He has worked extensively on quantitative analysis of energetic particle and magnetometer data from a range of space- and ground-based instrumentation, and on the validation of predictive models. Recent highlights include developing prototype realtime versions of the DREAM and SHIELDS

modeling systems, and development of the first ensemble prediction system for geomagnetic disturbances. He will contribute to the analysis and interpretation of space weather simulations and data.

Technology Transfer

1. Do you envision that this LDRD project might generate intellectual property (e.g. technology that can be patented or software that can be copyrighted)? (**yes**/no)

If so, please explain (100 word limit)

YES

We will be developing new statistical algorithms for analyzing large-scale data sets, as well as source code for embedding these statistical algorithms inside high performance computing codes, and a general programming environment for doing statistics on data as it is being generated by another software code.

2. Is there any encumbrance on the Laboratory's pursuit of that intellectual property (e.g. co-invention with other organizations, dependence on intellectual property that originates outside the Laboratory)? (yes/**no**)

NO

3. Will engagement with industrial partners enhance your technology and/or help you to realize the full possibility of this line of research and development? (**yes**/no)

YES